DATA-SPECIFIC CONCEPT CORRELATION ESTIMATION FOR VIDEO ANNOTATION REFINEMENT

Cencen Zhong, Zhenjiang Miao

Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China
E-mail: 07112072@bjtu.edu.cn, zjmiao@bjtu.edu.cn

ABSTRACT

For video annotation refinement, a reasonable concept correlation representation is crucial. In this paper, we present a data-specific concept correlation estimation procedure for this task, where the resulting correlation with respect to each data encodes both its visual and high-level characteristics. Specifically, this procedure comprises two major modules: concept correlation basis estimation and data-specific concept correlation calculation. Under the framework of sparse representation, the former introduces a set of high-level concept correlation bases to represent the concept distribution of each feature-level basis, while the latter constructs the concept correlation of a specific data by combining its feature-level sparse coefficients and correlation bases together. In the end, given this new correlation, a probability-calculation based video annotation refinement is performed on TRECVID 2006 dataset. The experiments show that such a representation capturing data-specific characteristics could achieve better performance, than the generic concept correlation applied to all data.

Index Terms— video annotation refinement, concept correlation, sparse representation

1. INTRODUCTION

Recently, for the sake of better performance, concept correlation defining the relationship between concepts has played an important role in video annotation (or concept detection) refinement [1,2,3]. Especially for the popular Context-based Concept Fusion (CBCF) [4], this contextual knowledge is encoded into a context-based model and acts as a post-processing to refine the initial results derived from individual concept detectors.

Considered as the guidance to refinement, a reasonable concept correlation representation is crucial. Normally, the statistics of concept co-occurrence and their extended forms such as normalized mutual information [3] and Pearson product-moment correlation are leveraged, producing a single concept correlation applied to all data. Although this generic correlation ideally carries exact relationships among all concepts, it may be not as effective as expected in practical video annotation refinement, e.g. in the scenario that ‘Sky’ and ‘Beach’ are strongly related in the context of outdoor while almost unrelated for meeting, the correlation between ‘Sky’ and ‘Beach’ should vary with the data to be refined other than keep constant. In other words, though this generic correlation has learned an 'ideal' relationship of them, it could not simultaneously meet both of these two cases, even neither of them, instead, becomes biased to hurt the refinement performance. Thus, we claim that in contrast to generic correlation, a finer representation dependent on specific data is expected to yield a better refinement result.

In other side, these aforementioned measures usually stem from a large-scale text dataset like the ground truth annotations, Word Net [6] and web-search lists [7]; and hence generate pure high-level generic concept correlations. However, in fact, except for this high-level textual clue, visual features regarded as the external representations of concept also contribute to concept correlation. This factor is emphasized in [7] by presenting a visual concept correlation based on visual similarities between images, but its linear summation with the high-level one still results in a generic relationship, having suppressed the visual property of specific data, e.g. visually similar data often contain the same concepts and accordingly may convey similar concept distributions. In this sense, a reasonable combination of visual characteristics into data-dependent correlation is also considered important for the following refinement.

Motivated by these observations, this paper presents a data-specific concept correlation estimation procedure for video annotation refinement, where the resulting correlation with respect to each data encodes both its visual and high-level characteristics in a natural way. The basic idea is as follows: under the framework of sparse representation, the original feature-level space could be broken into several clusters, each conveying certain ‘representative visual object’ and accordingly manifesting ‘representative concept distribution’; likewise to the effect of feature-level bases, these distributions are expected to be capable of spanning the high-level correlation space, and thus referred to as concept correlation bases in this paper; finally together with feature-level sparse coefficients these new bases are used to construct the concept correlation of a specific data, just analogous to the reversed process of data decomposition. To

this end, two major modules are involved in this work: concept correlation basis estimation and data-specific concept correlation calculation. Specifically, on basis of sparse representation, the former derives a set of concept correlation bases to represent the concept distribution of each feature-level basis, while the latter constructs the concept correlation of a specific data by combining its feature-level sparse coefficients and concept correlation bases together. Finally, unlike those complicated context-based models commonly used in CBCF, this paper formulates this new correlation into a probability calculation based scheme to refine the initial results derived from multiple detectors. And the final experiments conducted on TRECVID 2006 dataset demonstrate its effectiveness.

The rest of this paper is organized as follows: Section 2 presents the whole refinement framework; Section 3 details the key modules; and the experiment setup and results are described in Section 4; finally the conclusions are given.

2. DATA-SPECIFIC CONCEPT CORRELATION ESTIMATION FOR ANNOTATION REFINEMENT

In this paper, data-specific concept correlation based video annotation refinement is proposed as an instance of CBCF. As formulated by Eq.1, for a shot x_n classified by concept detector C_i ($i = 1, \ldots, J$, J concepts in total), its refined posterior probability $\tilde{p}(C_i|x_n)$ is a result of the interaction of initial result $p(C_i|x_n)$ and other detectors $p(C_j|x_n)$ weighted by data-specific concept correlation $p(C_i|C_j, x_n)$. Additionally, a tradeoff factor λ $(0 \le \lambda \le 1)$ is introduced to tune the contributions of initial result and refinement term.

$$\begin{aligned}\tilde{p}(C_i|x_n) &= \sum_{j=1}^{J} p(C_i, C_j|x_n) = \sum_{j=1}^{J} p(C_i|C_j, x_n)\, p(C_j|x_n) \\ &= \lambda p(C_i|x_n) + (1-\lambda)\sum_{j=1, j\neq i}^{J} p(C_i|C_j, x_n)\, p(C_j|x_n)\end{aligned} \quad (1)$$

Central to this formulation, $p(C_i|C_j, x_n)$ is calculated based upon the following two modules:

1. Concept correlation basis estimation: to derive Q feature-level bases d_q $(q = 1, \ldots, Q)$ and Q concept correlation bases $p(C_i|C_j, \Omega_q)$.
2. Data-specific concept correlation calculation: to derive data-specific correlation $p(C_i|C_j, x_n)$ utilizing d_q and $p(C_i|C_j, \Omega_q)$ obtained above.

So far, the final refinement algorithm could be rewritten as Eq.2, where $\alpha_{n,q}$ is the sparse coefficient for shot x_n.

$$\tilde{p}(C_i|x_n) = \lambda p(C_i|x_n) + (1-\lambda)\sum_{j=1, j\neq i}^{J}\left(\sum_{q=1}^{Q}\alpha_{n,q} p(C_i|C_j, \Omega_q)\right) p(C_j|x_n) \quad (2)$$

3. KEY MODULES

3.1. Concept correlation representation

In the light of the probability-calculation based refinement described by Eq.2, the conditional probability that concept C_i appears with the existence of C_j is adopted to represent the concept correlation in this paper. Given the concept co-occurrence information in a ground truth annotation set Θ, it is calculated as:

$$p(C_i|C_j, \Theta) = \frac{Count(C_i, C_j|\Theta)}{Count(C_j|\Theta)} \quad (3)$$

where $Count(C_i, C_j|\Theta)$ is the count of C_i and C_j concur within the same shot belonging to Θ. $Count(C_j|\Theta)$ is the count that C_j occurs in Θ.

3.2. Concept correlation basis estimation

Sparse representation, decomposing signals into a series of sparse coefficients with respect to fixed bases, has been well applied in many fields, such as face recognition, image denoising, etc [8,9]. In the same way, each video shot also could be decomposed into some fixed feature-level bases, along with corresponding coefficients. Starting from this, for each feature-level basis, those shots that have made efforts to its generation are likely to convey some common visual characteristics and thus clustered together to denote implicit 'representative visual object'. After that, the concept correlation gathered within each cluster results in one correlation basis to describe the concept distribution of each feature-level basis, and meanwhile is assumed to be representative enough to span the correlation space, just similar to the effect of feature-level bases. To determine these new bases, the detailed steps are given as follows:

Step.1: For the training dataset, the dictionary learning and sparse representation algorithm is employed [10], to obtain the dictionary D and the sparse coefficient β_m with respect to each data x_m $(m = 1, \ldots, M)$.

$$\min_{D\in\Delta, \beta\in^{Q\times M}} \sum_{m=1}^{M}\frac{1}{2}\|x_m - D\beta_m\|_2^2 + \gamma\|\beta_m\|_1 \quad (4)$$
$$\Delta = \left\{D \in^{dim\times Q}, s.t. \forall q = 1, \ldots, Q, \|d_q\|_2 \le 1\right\}$$

where γ is the sparsity measure. dim denotes the dimension of feature space and Q is the number of feature-level bases constituting the dictionary D.

Step.2: By taking the learned dictionary as cluster centers, all original data are reorganized into clusters, i.e. for each feature-level basis d_q, those data whose coefficients satisfie $\beta_{m,q} > 0$ are assigned to cluster Ω_q.

Step.3: Grounded on Eq.3, the concept correlation bases $p(C_i|C_j, \Omega_q)$ are estimated within each cluster Ω_q.

3.3. Data-specific concept correlation calculation

In the reversed process of data decomposition, the derived feature-level bases together with sparse coefficients could reconstruct the original shot. In like manner, it is supposed that concept correlation bases spanning the correlation space

could also form the data-specific concept correlation, via treating those feature-level coefficients as weighting factors. In other words, by acting like that, this procedure provides a natural solution to combine the data-specific visual (sparse coefficients) and high-level (concept correlation basis) characteristics together. In detail, the calculation for $p(C_i|C_j, x_n)$ is described like:

Step.1: For each testing shot $x_n, (n = 1, \ldots, N)$, it is sparsely represented by D according to Eq.5.

$$\min_{\alpha \in \mathbb{Q} \times N} \sum_{n=1}^{N} \frac{1}{2} \|x_n - D\alpha_n\|_2^2 + \gamma \|\alpha_n\|_1 \quad (5)$$

Step.2: The coefficients $\alpha_{n,q}$ are leveraged as the weighting factors to combine the concept correlation bases $p(C_i|C_j, \Omega_q)$ together to form the final $p(C_i|C_j, x_n)$

$$p(C_i|C_j, x_n) = \sum_{q=1}^{Q} \alpha_{n,q} p(C_i|C_j, \Omega_q) \quad (6)$$

Finally, according to Eq.1, the refinement is carried out based on $p(C_i|C_j, x_n)$ and posterior probabilities of all detectors $p(C_j|x_n)$ including $p(C_i|x_n)$.

4. EXPERIMENT

4.1. Experimental setup

In this paper, we adopt TRECVID 2005 dataset including its ground truth annotations for concept correlation basis estimation, and TRECVID 2006 for evaluation on data-specific correlation based refinement. Since we only focus on the result of detectors other than their construction, 374 concept detectors provided by Columbia374 are employed as baseline [11]. In accord with the test setting in TRECVID 2006, we only utilize 20 official concepts in light evaluation to evaluate the results, in terms of inferred average precision (infAP) for each concept and mean infAP (MAP) over all concepts. Note that the outcome of Columbia374 on TRECVID 2006 testing data, adopted as the initial result to be refined later, has gained an MAP of **0.0948**.

4.2. Experimental result and discussion

This section details our experimental results with different parameter settings. In addition, a detailed comparison on generic and data-specific concept correlations is given as well.

4.2.1. Experiment 1: generic concept correlation based video annotation refinement

As an approximation of $p(C_i|C_j, x_n)$, the generic concept correlation $p(C_i|C_j)$ estimated over all annotations is utilized to conduct the refinement according to Eq.1. To balance the contributions made by initial result and refinement term, we list its refinement results under different tradeoff factor λ settings.

Table.1. Performance on generic concept correlation based refinement with different tradeoff factors

λ	0.1	0.15	0.25	0.35	0.5	0.75
MAP	0.1190	**0.1199**	0.1176	0.1149	0.1083	0.1002

Table.1 tells that MAPs at different λ are all larger than that of baseline, which has validated the positive role of concept correlation in annotation refinement. According to these results, we select the optimal $\lambda = 0.15$ for our proposed procedure in next subsection.

4.2.2. Experiment 2: data-specific concept correlation based video annotation refinement

As it concerns sparse representation, we empirically let the sparsity measure γ in Eq.4 and Eq.5 be 4. Given $\lambda = 0.15$ determined in Section 4.2.2, Table.2 lists the performance of our method with different numbers of feature-level bases.

Table.2. Performance on data-specific concept correlation based refinement with different numbers of feature-level bases

Q	10	20	30	40	50
MAP	0.1246	0.1266	0.1276	0.1282	0.1293
Q	60	70	80	90	100
MAP	**0.1305**	0.1301	0.1293	0.1289	0.1273

As shown in Tabel.2, when $Q = 60$ our approach gains the best MAP, with around 37.66% improvement compared to baseline, and 8.84% to generic concept correlation based refinement. Such a promising result has demonstrated that correlation bases which decompose the correlation space do have made contributes to enhancing the descriptive ability of concept correlation and thus devoted efforts to improve the refinement. On top of this, it is also found that when $Q < 60$ the performance is getting better with the increase of number of feature-level bases. This is consistent with the conclusion above that finer correlations are more likely to denote these various 'representative objects' and hence could intensively capture the data content. While for $Q > 60$, MAP has decreased but is still larger than that of generic one. We attribute this slight drop to the over fitting correlation bases, which are caused by the limited size of ground truth annotation set.

4.2.3. Comparison on generic and data-specific concept correlations

In addition to the comparison on MAP, this subsection gives an intuitive insight into concept distributions supplied by generic and data-specific correlations. We take a testing shot involving 'Police_Security', 'Military', 'Car', 'Truck' and 'Explosion_Fire' (they are termed as related concepts, while the other 369 as unrelated) as an example. For simplicity, we only exemplify the correlations between 'Police_Security' and 35 concepts in Fig.1, where the brighter the block is, the stronger the correlation is.

It can be seen that, for data-specific correlation, the distribution of related concepts is more or less consistent

with the generic one, which has capsulated the major data content. But better than that, in data-specific representation the correlations between 'Police_Security' and some unrelated concepts like 'Outdoor', 'People-Marching' and 'Face' are suppressed as well, enabling it to reduce the risk of bringing in unnecessary noise. Thus from this perspective, we can say that the data-specific correlation could be more coherent with the data content and consequently lead to a better refinement performance. In addition, it is worth to note that the correlation between 'Police_Security' and 'Prisoner' is reinforced. Indeed, among the limited data annotated by 'Prisoner' (133 shots) about 25% (30 shots) are simultaneously labeled by 'Police_Security'. Therefore, it is easy to assign the shots that contain 'Police_Security' as well as other concepts related to 'Prisoner' with large coefficients with respect to feature-level bases dominated by 'Police_Security' and 'Prisoner'. Even so, whether the final result involves the unexpected 'Prisoner' or not also depends on the initial detection scores. But still, it is reasonable to suppose that under a larger-scale working set the proposed data-specific concept correlation would work better.

5. CONCLUSIONS

In this paper, we have proposed a data-specific concept correlation estimation procedure for video annotation refinement. Firstly, a set of concept correlation bases is introduced to capture the representative high-level concept correlation corresponding to each feature-level basis. Then, for a specific data, its concept correlation is calculated like a reversed process of data decomposition, based on feature-level sparse coefficients and correlation bases. This data-specific correlation encoding both visual and high-level characteristics has been proved effective on TRECVID 2006 testing set. And compared to the generic concept correlation, our approach could achieve better performance.

In the future, as concluded in Section 4.2.3, a larger-scale working set, such as web-search lists, will be adopted to generate a richer representation. Furthermore, allowing for the content evolvement in video stream, temporal concept correlations will be considered as well.

ACKNOWLEDGEMENTS

This work is supported by NSFC 60973061, 973 Program 2011CB302203, and Ph.D. Programs Foundation of Ministry of Education of China (20100009110004).

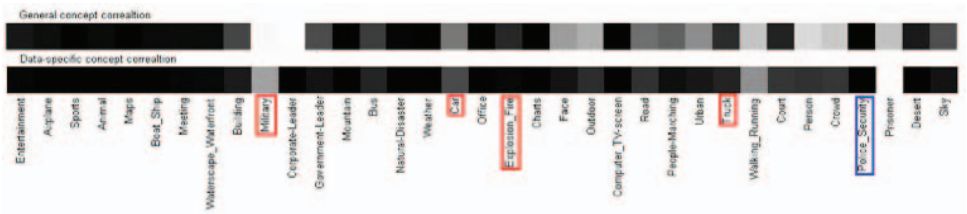

Figure.1. An example of comparison on generic (upper part) and data-specific concept correlation (lower part), for testing shot ''shot12_207' between 'Police_Security' and other 34 concepts truncated from the original 374-concept set

6. REFERENCES

[1] C. G. M. Snoek, M.Worring, J. C. van Gemert, J.M. Geusebroek, and A. W. M. Smeulders, "The Challenge Problem for Automated Detection of 101 Semantic Concepts in Multimedia" , in *ACM Multimedia*, USA, pp.421-430, 2006.

[2] Y. Li, Y. Tian, L. Duan, J. Yang, T. Huang, and W. Gao, "Sequence Multi-Labeling: A Unified Video Annotation Scheme with Spatial and Temporal Context", *IEEE Trans. Multimedia*, pp. 814-828, 2010.

[3] G. Qi, X. Hua, Y. Rui, J. Tang, T. Mei, M. Wang, and H. Zhang, "Correlative Multilabel Video Annotation with Temporal Kernels" , *ACM Trans. Multimedia Comp.,Comm.,and Appl.*, pp.1-27, 2008.

[4] M.R. Naphade, I. Kozintsev, and T.S. Huang, "Factor Graph Framework for Semantic Video Indexing", *IEEE Trans. Circuits Syst. Video Techn.*, pp.40-52, 2002.

[5] X. Wei, Y. Jiang, and C. Ngo, "Exploring Inter-concept Relationship with Context Space for Semantic Video Indexing", in *ACM Image and Video Retrieval*, Greece, 2009.

[6] A. Torralba, R. Fergus, and W.T. Freeman, "80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp.1958-1970, 2008.

[7] J. Liu, M. Li, Q. Liu, H. Lu, and S. Ma, "Image Annotation via Graph Learning," *Pattern Recognition*, pp.218-228, 2009.

[8] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Y. Ma, "Robust Face Recognition via Sparse Representation," *IEEE Trans. Pattern Anal. Mach. Intell*., pp.210-227, 2009.

[9] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Non-local Sparse Models for Image Restoration," in *ICCV*, Japan, pp.2272-2279, 2009.

[10] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online Dictionary Learning for Sparse Coding," in *ICML*, Canada, 2009

[11] A. Yanagawa, S.F. Chang, L. Kennedy, and W. Hsu, "Columbia University's Baseline Detectors for 374 LSCOM Semantic Visual Concepts," *Columbia University ADVENT Technical Report #222-2006-8*, March 20, 2007.